Towards a semantic data library for the social sciences

Grotton, Thomas; Hachenberg, Christian; Harth, Andreas; Zapilko, Benjamin

Towards a Semantic Data Library for the Social Sciences

Thomas Gottron[1], Christian Hachenberg[1], Andreas Harth[2] and Benjamin Zapilko[3]

[1] WeST – Institute for Web Science and Technologies, University of Koblenz-Landau, Koblenz, Germany
[2] Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, Germany
[3] GESIS – Leibniz Institute for the Social Sciences, Knowledge Technologies for the Social Sciences, Bonn, Germany
{gottron, hachenberg}@uni-koblenz.de, harth@kit.edu, benjamin.zapilko@gesis.org

Abstract. Quantitative research in the Social Sciences heavily relies on survey and statistical data. While researchers often put a lot of effort in generating such data, the incorporation and reuse of existing data on the web is far behind its potential. The lack of reuse can be attributed to various deficits in terms of library services, in particular, common exchange formats, annotations with metadata or standard approaches for integrating and merging data sets as well as the lack of an easy approach for searching data records and a lack of publicly available data sets. To overcome such problems which in the past have already been addressed by libraries, we propose a framework for seeking, merging, integrating and aggregating distributed statistical and survey data based on open semantic formats. We present a first prototype implementation as show case for the framework and highlight the benefits for social scientists.

Keywords: Semantic Digital Data Library, Linked Data, Statistics, Data Integration

1 Introduction

Libraries and archives follow a long tradition in surveying, collecting and classifying available knowledge and in providing access to these high quality information resources. With the distributed publishing paradigm of the web, providing such services has grown in complexity, driven by a multiplicity of exchange formats, different terminologies for metadata annotations and missing connections between distributed data sets. However, researchers cannot use distributed data on the web in the same way as they are used to in libraries and archives. One reason is that Digital Libraries and Digital Archives are often still disconnected from each other – not only because of historical and disciplinary reasons, but also because they use different standards and formats.

Research in the Social Sciences often relies on empirical data for studies. The emerging field of "Computational Social Sciences" leverages the possibility of collecting and analysing large-scale datasets to potentially reveal patterns of behaviour of individuals and groups [16]. The necessary data for such an approach is often difficult to find, integrate and process, which is due to a mostly decentralised

and historically grown distributed publication and archiving of data in e.g., government agencies, research data centres or universities. Scattered information due to organic growth also occurs on the web at large. To be able to judge the relevance and quality of the data for any upcoming analysis in research, it is important to gain deep insights into both data and especially its documentation. Besides descriptive standard information, the metadata of data used in analysis shall provide extensive information about methodology, sample design, necessary weights or notes on the safe and correct handling of the data concerning privacy and provenance. A lack of metadata annotation complicates the process of data search on the web as well as the comparison of different data sets, e.g., regarding concrete indicators or populations.

While sizable amounts of data useful for research are attainable through the web, the data is published in a large variety of data formats. To process and analyse data, one has to convert data into particular formats of statistic tools, and integrate data from multiple sources. In general, data conversion and integration is not a technical barrier, but the effort spent for conversion is a nuisance, especially for necessary but tedious routine tasks, such as gaining a first insight into the data, or in cases where the expected research gain is minor. All these problems hinder a reuse of available and valuable data resources.

To overcome the challenges that Digital Libraries and Archives are facing with distributed data on the web, we propose a framework for a Semantic Digital Library of Linked Data, which is relevant for research in the Social Sciences. While the framework provides central services for accessing, processing and integration of distributed data sources, their physical storage location remains distributed and will not be collected or hosted by the data library. The difficulties in searching, modelling and annotating distributed data are addressed not only on the metadata level, but also on the directly connected underlying numerical data, which provides researchers an on-the-fly usage of the data in visualisations or for statistical analysis. We present a prototype implementation which demonstrates the automatic aggregation and integration of data using wrappers and a common exchange format.

The rest of the paper is structured as follows. In Section 2 we present a use case of a typical research scenario in the Social Sciences. Section 3 provides related work regarding Linked Data and the use of semantic technologies for processing data. We present existing data formats for modelling statistical and survey data in Section 4. In Section 5 we propose a framework of key modules for a Semantic Digital Data Library. Results of a first prototype implementation are presented in Section 6 and open issues are discussed in Section 7. We conclude and present future work in section 8.

2 GESIS Use Case

As an organisation providing infrastructure for the Social Sciences, GESIS – Leibniz Institute for the Social Sciences[1] offers a wide range of different study series as well as empirical primary data from survey research and historical social research. At the beginning of any research, scientists usually have a first idea what kind of data they

[1] http://www.gesis.org/

will need and which analysis method they would like to perform on the data. For example, a researcher would like to investigate possible correlations in a correspondence analysis of unemployment rate, immigration quota and the subjectively perceived risk of unemployment in Germany. However, the desired data is only available from different authorities. While the researcher can retrieve statistics from German statistical offices, data on attitudes, behaviour and social structure in Germany is part of the German General Social Survey ALLBUS[2], which is archived at GESIS. On the web portals of GESIS, the ALLBUS metadata can be searched, so the researcher can gain insight into the documentation of the data and is able to decide, whether ALLBUS is (completely or partly) relevant to the research interests. For a decision, whether the data is suitable for the intended analysis method, a comprehensive and detailed documentation of the data is essential. Information on e.g., sample design, populations or possible bias and variance has to be provided. In case researchers would like to analyse more than one data set, the individual data sets have to be aligned, i.e., not only technically, but also considering differences in populations or aggregation levels.

Using statistics tools such as STATA[3], SPSS[4] or the R Project[5] might require the data to be converted into application-specific formats. When dealing with different data sets, it has to be clear what dimensions and samples the data is comparable to and thus how data can be matched up. For example, data from ALLBUS has to be aggregated to be comparable to any statistics, because ALLBUS is micro data and therefore determined at individual level due to its origin as survey data. The matching is mostly done manually before importing the integrated data into statistics tools, although some tools can automatically detect comparable dimensions like time or geographic regions. Finally, the research analyses data and defines and executes statistical functions, which depend on the desired analysis method such as multidimensional analysis, time series analysis, correspondence analysis or estimation procedures in complex designs [12][13][17].

After finishing research and data analysis, researchers ought to cite the used data sets in the resulting publications. Referencing the analysed data helps fellow researchers to comprehend the analysis done with the data. Data can be cited and afterwards identified by using a URI (Uniform Resource Identifier) or a DOI (Digital Object Identifier). Newly created data during research obtains an identifier only if it is published afterwards.

3 Related Work

Semantic Data Libraries and Archives address key challenges like information integration and interoperability as well as user-friendly interfaces, all supported by semantic technologies and community interactions [14]. They are the next step and further evolution of traditional digital approaches, which often lack the implementation of Semantic Web and social networking technologies. Considering a

[2] http://www.gesis.org/en/allbus
[3] http://www.stata.com/
[4] http://www.spss.com/
[5] http://www.r-project.org/

Digital Library of distributed data, semantic technologies can facilitate the integration of data from disparate sources.

In recent years the idea of Linked Open Data [3] emerged. Linked Open Data represents a way to expose, share and connect freely available data on the web using Semantic Web standards. The publication of data as Linked Open Data from a technical perspective [2] is based on common standards and techniques which have been developed for years and are established worldwide as fundamental formats and interfaces for publishing data on the web, e.g., URIs, HTTP and RDF. With the standardisation of SPARQL [19], a common technology for querying RDF data has been established. The paradigm of Linked Open Data was well received in the Semantic Web community and has encouraged organisations worldwide to publish data. In recent years a lot of statistics and other numerical data have been published as Linked Data by e.g., government agencies, statistical offices or research organisations. To find available data sources, open data repositories like the Data Hub[6] have been established, where data sets can be described and grouped. Currently a common vocabulary, the Data Catalog Vocabulary[7], for the description of such data sets is under development.

Semantic technologies can aid in the integration and combined querying of data. Both descriptions of a data set (such as author, publication date) and the data set itself (individual observations) can be encoded and interpreted by machines. Thus, the integration is made possible. We present different data formats in the next section in more detail. Both are required: descriptions of the data (e.g., author, responsible organisation) and the data itself (the individual observations). Once data has been published in a uniform base format (e.g., RDF), machine-supported integration is possible. There are several services possible on integrated data, for example keyword search [15] or faceted browsing [22]. VisiNav in particular offers navigation functionality over data integrated from the Web [10]. OLAP clients may be used to perform analysis queries on the integrated data. An overview on semantic web search is given in [21]. Another way to query the data is via SPARQL. The SPARQL plugin for the R Project[8] - an open source software environment for statistical computing - allows for the formulation of SPARQL queries within R and the use of the retrieved Linked Data for statistical calculations.

Retrieving and analysing data on the web is nothing new to researchers in the Social Sciences. Data providers of statistical or survey data are very keen on offering the possibility for browsing, analysing and downloading their data, even if it is only metadata due to privacy restrictions. Examples are ZACAT[9] and SOEPinfo[10]. Both portals offer a wide range of tools for processing, analysing, the visualisation and export of data to different data formats. However, both are restricted to the data holdings of their particular organisation. A web-based application which is more open is GraphPad QuickCalcs[11], a collection of free online services for e.g., statistical calculations based on data manually entered by the user. However, calculations are

[6] http://ckan.net/
[7] http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary/Vocabulary_Reference
[8] http://cran.r-project.org/web/packages/SPARQL/
[9] http://zacat.gesis.org/
[10] http://panel.gsoep.de/soepinfo/
[11] http://www.graphpad.com/quickcalcs/index.cfm

only possible on single numbers and not on entire data sets. As yet, such data analysis tools are not empowered by semantic technologies. However, [9] identify large potential impact in the use of such technologies and available Linked Data for research activities in the Social Sciences.

4 Data Format for Statistical Data

When considering Data Library services for statistical and survey data, the proper format to store and exchange/transform data is a key component. In this section we present the formats which are most relevant to our task.

SDMX (Statistical Data and Metadata Exchange) [20] was established in 2002 by key players in the field of statistical data, such as the World Bank, IMF and the European Central Bank. Paramount was the ability to enable automatic machine-to-machine exchange of data, which requires a self-expressive or self-descriptive metadata model. SDMX defines representations of statistical data and respective metadata annotations, not only for single data items but also for full data sets. The SDMX information model is based on named concepts which are assigned *dimensions* and *attributes*. Dimensions can be grouped into so-called *keys* using *code lists* for available realisations; plain free-text is allowed as well. *Data Structure Definitions* assemble all these components with respect to a specific topic or data source in a well-defined structure. In this way, multidimensional statistical data can be represented by the SDMX information model. As we will elaborate below, parts of SDMX are reused in the definition of the Data Cube metadata model.

SCOVO (Statistical Core Vocabulary) [11] is an RDF-Schema based, lightweight vocabulary for representing statistical data. As such, SCOVO aims for an eased community uptake (since statistical data formats in general are rather complex to use) and promotes the Linked Data publishing principles, which on the one hand require use of RDF and on the other hand include re-usage of existing and well-established vocabularies, such as SKOS (Simple Knowledge Organization System). SCOVO thus fosters extensions both on the schema and instance level. Another important design issue for SCOVO was – in line with SDMX features – the ability to handle as many dimensions as necessary (supporting a multidimensional model). Compared to SDMX's focus on generic and efficient data exchange, SCOVO has weaknesses under this aspect. Being part of the Web of Data and complying to RDF standards as message format enables both self-descriptive data items and generic data exchange. SCOVO consists of mainly three principal classes: *item, dimension* and *dataset*. The first describes a single observation or event. The second describes and identifies the contents of an item whereas the latter is made up of a number of items sometimes, also defining a concept (which is provided as SKOS concept).

The RDF **Data Cube** (QB) vocabulary and its metadata model [5] is another way of representing multidimensional statistical data in RDF following the Linked Data principles (and can be seen as successor of SCOVO). To date, the vocabulary still exists only as a draft but is supposed to become widely accepted in the future due to its various advantages (see also paragraph 5.2). In particular, QB incorporates all the features of SCOVO but goes beyond some of its limitations. The Data Cube vocabulary makes use of relevant parts of the SDMX information model. For the RDF

part, QB can use language descriptors of SKOS [18], FOAF [8], VoiD [1] and Dublin Core terms [7]. The metadata model of Data Cube implements the idea of a multidimensional „cube“ where all data points (i.e., observations) are aligned along certain edges and one can cut „slices“ through the cube to get cross-section and low-dimensional data views. QB also has components like *dimension*, *measure* and *attribute* which are all set up in a *data structure definition* class. The semantics of dimensions and attributes are similar to SCOVO or SDMX. Dimensions describe what is observed when considering a single data item whereas a measure describes the overall phenomenon being measured or represented for a single observation. Statistical concepts can also be defined and assigned to a SKOS concept, similar to SCOVO. Furthermore, one can add metadata to data sets using Dublin Core terms or to single observations using the attribute component. Observations are organised in data sets and hold the actual values which are categorized by dimension, measure and attribute, in turn. According to [5] Data Cube is unique in its features compared to SCOVO.

In contrast to aggregated data, so-called micro data in the Social Sciences is described by the **DDI (Data Document Initiative)** [6] metadata specification, which is an international standard describing and maintaining survey data in the social, behavioural and economic sciences. One of the key features of the DDI format is the documentation of the entire research data life cycle, which includes activities on data from the conceptualisation, collection and processing of survey data to their analysis and archiving. The complexity of DDI enables the possibility to document data very extensively, which is necessary for researchers to search and judge data according to relevance and quality. Because micro data is an important basic source for aggregated data, there are crucial similarities and overlaps. However, existing mappings are often undocumented. Since 2009, a working group is defining a detailed mapping between DDI and SDMX. Until now, there is no representation of DDI in RDF, but the process of designing a DDI ontology has begun [4].

5 A Framework for a Semantic Library of Statistical Data

To address the key challenges for semantic library services for survey and statistical data in the Social Sciences, we introduce a generic framework. The framework is composed of modules for identifying and exchanging, searching and integrating, evaluating and publishing data. Thereby we address the main obstacles for reusing statistical or survey data in the Social Sciences, also related to our GESIS use case.

5.1 Common Identifier Format

Identification of data sets, measurements or dimensions is of importance for a variety of reasons. On the data level a unique identifier allows for referencing the data set itself. Referencing is crucial in the context of making data sets citeable in scientific publications, thereby providing valuable metadata about the scientific work. Within the data, the identifiers provide a way to identify the semantics of dimensions, measures and observations. URIs fulfil this requirement and are a core ingredient to

semantic web technologies. With respect to integration and aggregation of data sets, in particular the semantics of the dimensions is of interest.

5.2 Common Exchange Format

There are a couple of well-established and proven formats for statistical calculations. Amongst others, Excel spreadsheets, SPSS, SAS, Stata or R native formats are used to carry around data including respective formulas. Unfortunately, these formats are proprietary (locked) and/or in binary format, which makes it difficult to transform data seamlessly from one format to another. Additionally, all these well-known formats do not describe their data in an expressive way, i.e., expressive enough to deliver self-explanatory data via metadata. For the purpose of a data library for the Social Sciences, it is necessary to integrate various heterogeneous data sources and perform calculations directly on data or on aggregated items coming from these sources. To achieve direct calculations, we are interested in self-explanatory or self-descriptive data sources which deliver generic structures which can be semantically processed further on. Thus, we aim for annotated or metadata-enriched data formats which promote easy exchange, integration and annotation using data from many, heterogeneous sources. These requirements are well met by the Data Cube format since it is (a) an open, non-proprietary metadata model in RDF format, (b) widely based on the established SDMX information model and also including other vocabularies, (c) provides a semantic and self-descriptive annotation of the data. Given these advantages it is likely that this metadata model will be supported by established statistics packages or that converter programs will be developed. The advantages of QB foster a thorough adoption by practitioners and facilitate an easy deployment and publication of statistical and survey data. Another advantage of Data Cube is that thanks to its flexibility and simplicity it is easy to convert existing data. In our prototype implementation presented below, we actually use efficient wrapper modules to convert proprietary or other non-semantic formats on-the-fly to the Data Cube vocabulary.

5.3 Retrieval of Statistical Data

The ability to find relevant data sets is a key factor to enable social scientists to make use of existing data sets. Therefore an efficient retrieval module is necessary for search of data being suitable for the respective research topic. Later on in the retrieval process more details about the requested data become evident, for example the granularity of specific dimensions or the frequency of observations. To provide researchers with useful information about a data set, there has to be extensive metadata available. Metadata not only supports the retrieval process itself, but has also to be considered afterwards to be able to evaluate relevance, quality and suitability for the following analysis process. For comparative research the description and attributes of for example different indicators, sample designs and populations have to allow for comparisons to those of other data sets. Eventually, the retrieval module should provide the underlying data itself.

The semantic description of the data also enables more complex search tasks. For instance, if a researcher is interested in the GDPs of European countries, the available

data provides these figures in the currency of the corresponding countries and not all of the data might be provided using Euro as a currency. If a second source can deliver the conversion rate, it is possible to combine the data sets and produce the requested information. Beyond the actual retrieval of the data sets, the module will need to provide a simple interaction component to define possible common dimensions by which data sets should flexibly be merged and integrated, i.e., time or geographical areas. Therefore the task of the retrieval module is twofold: retrieve (a) metadata about the datasets (e.g., using taxonomies, as common in libraries - SKOS) and (b) the data sets themselves.

5.4 Data Linking and Integration

The semantic representation and annotation of data allows for services far beyond the simple retrieval and provisioning of data sets. As the semantics of dimensions, values and metrics is explicitly modelled in the data, automatic linking and integration of data is at a researcher´s fingertips.

To correctly join and merge two data sets it is necessary to identify common dimensions, align and map the according values and possibly aggregate some of the data entries. Based on the dimension concept in Data Cube and the possibility for semantic annotation, the identification step can be made without any efforts. Alignment of the values requires some more insights and may be achieved by a more detailed model and description of the data. On data with temporal dimension, for instance, it is necessary to define its resolution and differentiate between hourly, daily, monthly, quarterly or yearly values. Aggregation becomes necessary when there is no direct representation and the data values need to be summed, or averaged. Again the semantic description of the dimension may provide exactly the information necessary to know which aggregation function to apply.

5.5 Preview and Analysis

For any existing or newly created (by the means of linking and integration) data set, the first approach for a social scientist will typically be to take a look at some key characteristics of the data. Therefore, together with the provision of the data itself, the library will present some results of a simple statistical analysis. For existing data sets key characteristics can be pre-computed, for freshly integrated data an overview will be generated on-the-fly. Once more, we benefit from a semantic representation of the data that allows for a better notion of which characteristics will be of interest and which dimensions need to be looked at.

To make an analysis at first glance even easier, data sets should be presented in a graphical form, plotting key indicators over the main or common dimensions of integrated data sets.

5.6 Data Export and Referencing

While the preview and basic analysis can provide first insights into the data it neither can nor is supposed to replace the analysis based on a full statistics application.

Therefore the system needs to allow for exporting the data to enable downstream processing. An export service providing data sets in a selection of common formats (like CSV, Data Cube, or Excel) is crucial to feed into the individual scientific processing pipelines of research groups. Exporters are needed in particular as long as the Data Cube format itself is not supported by all major statistics tools.

As each dataset is compiled based on user-defined parameters and needs, the dataset can be reproduced at any time. Parameters can also be used in a unique identifier to a data set. Thereby data sets can be referenced and cited.

6 Prototype Implementation

The motivation behind our prototype is to investigate further areas of research utilising state-of-the-art technologies. However, we keep the focus on integration and analysis of data since search/retrieval of data on the Semantic Web is an already established field of research. To identify data items and corresponding dimensions, measures or attributes, we use RDF URIs common to the Semantic Web together with data structures defined in Data Cube vocabulary. Data Cube compliant data is generated by on-the-fly wrappers from our IT.NRW data source and by a conversion of data exported from the ALLBUS database. We do not include search capabilities in our prototype for retrieval of data sets since we only process a few data sets. We therefore enable the user to manually select the data sets to be used from a fixed list. For the integration step, all data in Data Cube format is then collected in an RDF memory store and accessed via a SPARQL end-point on top of the RDF store. In our case, we use OpenRDF's Sesame[12] library including a SPARQL interface since the prototype is implemented as a Java-based web application on an Apache Tomcat infrastructure using servlets.

The concrete task for the prototype is to integrate, aggregate and visualize data from two sources, ALLBUS and IT.NRW[13], which has to answer the (exemplary) question to find correlations between the number of votes per party and the people's ratings of economic situation (both personal and national prospect) in the German state of North Rhine-Westphalia. Hereby, ALLBUS provides survey data of individuals rating personal and national economic situation. IT.NRW provides the number of votes per party of elections to the "Bundestag" (the German national parliament) for the state of North Rhine-Westphalia. Figure 1 provides an overview of the architecture for the prototype which can be accessed online[14]. During the implementation phase we came across challenges regarding aggregation using current technologies. Since we use SPARQL 1.0 for querying, aggregation on the query level is not possible (yet) due to lack of functionality in the SPARQL language. Aggregation has to be done on application level or data modelling level. For ALLBUS data we solely aggregate on the data level. We intended to use numbers for the whole state North Rhine-Westphalia, but since we only had data from individuals

[12] http://www.openrdf.org/
[13] http://www.it.nrw.de
[14] http://lod.gesis.org/gesis-lod-pilot/ (note: German only)

from that state we did an upscaling to the whole population. Such processes can be included into metadata in order to reproduce changes on the data.

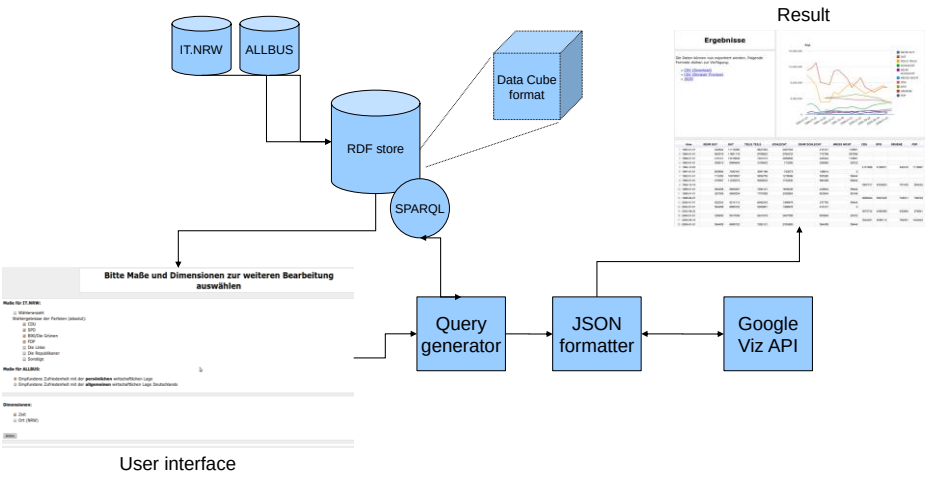

Fig. 1. Overview of implemented architecture

Analysis is both done visually and using lightweight calculations on integrated data. For the visualisation, we use the 2D line chart and table component from Google Visualization API which takes data in JSON format. So we transform SPARQL results to JSON just for displaying. Our visualisation allows for time-series analysis of election results in comparison to people's future prospects by analysing line charts or table data. For an experimental implementation of two statistical methods, calculations of variance and linear regression were integrated [23] on data coming from ALLBUS and Eurostat[15]. Both calculations are performed in Java since SPARQL does not provide for calculations yet. Eventually, data can be seamlessly exported to CSV and JSON for further analysis in e.g., external statistics tools.

7 Open Issues

There are several open issues in the realisation of a large scale Semantic Data Library for the Social Sciences. Some of which are of technical nature on a higher level (relative to the technical details identified in the prototype implementation), others are more related to the research culture of the potential user community.

One rather technical issue is how to deal with privacy. Survey data is anonymised to ensure the privacy of the participants. When merging and integrating data sets these anonymisation efforts can be annulled, as the combination of information allows for identification of individuals. To avoid such problems it is necessary to formalise,

[15] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/ via http://estatwrap.ontologycentral.com/

model and describe implications on the kind and type of data sets another data set may be combined and integrated with.

A similar meta-information that is crucial to a valid scientific analysis is the description of any bias present in the data. Statistical data is based on a sample of a larger population. The initial producers of such a data set are typically aware of any sampling bias they might have in the data (over- and underrepresentation of age groups, geographic location, cultural background, etc.). When publishing a data set on a library the knowledge of any bias needs to be preserved, which is of particular importance in a scenario where data sets are integrated and joined, as skewed bias may lead to wrong conclusions (e.g. joining data on perceived job-security and preferences of political parties sampled from different income groups).

To adequately address the issue with biased data as well as to enable (semi-) automatic merging, aggregation and integration of different data sources it is possibly necessary to further extend existing metadata models like Data Cube and/or complement with other vocabularies specifically dealing with data transformation. Bias in statistical data or other limitations of the data in use should have standardised support in terms of vocabulary in metadata models (e.g. descriptive comments are currently supported but lack the advantage of standardized vocabulary for automatic processing). However, more automatic data merging or aggregation needs standardised ways of applying transformation rules to deal with heterogeneous data structure. Here, specific vocabularies/ontologies for data transformation come into play, which is an open research issue.

A less technical issue is rooted in the scientific culture of the Social Sciences. The preparation and curation of data sets is a labour-intensive and time-consuming task. The work invested pays off in the production of high quality papers and an according reward in the sense of scientific reputation in the form of citations. Publishing a data set itself does not create citations (as there is no established process), and thus no scientific reputation. Therefore, data sets are rarely published, as data publication might actually bear the risk that other research groups come up with important findings quicker and thereby exploit the development of the data set without repaying the original work. While this behaviour is a cultural issue in the community of the Social Sciences, a Semantic Data Library which supports citation of data sets might have an impact on the behaviour. If a data set can be cited and thereby provide the authors with scientific credits, they might be less reluctant to publish their data. An issue related to citing data sets is the question of granularity. URIs actually allow for the “deep linking” of individual observations. How to enable fine-grained linkage and referencing with DOIs is an open question.

8 Conclusions and Future Work

We have presented a use case and associated requirements analysis for the publication of and integrated access to data relevant to research in the Social Sciences. During our analysis and discussions with social scientists we have identified the problem of locating relevant data sets, which has to be addressed before more elaborate integration and analysis functionality can be provided. We have presented a prototype implementation of a Semantic Data Library, which differs conceptually from

traditional libraries due to a publishing and integration process based on distributed Linked Data. The proposed framework covers the entire life cycle from publication to accessing data via software applications and a web application. Future work includes the addition of more sources to the data collection, better ways for establishing and using mappings between the different data sets, and a live deployment and evaluation of the approach with domain experts. Finally we plan on making the service publicly available on the web to start creating a community around public survey and statistical data sets in the Social Sciences.

9 References

1. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing Linked Datasets with the VoID Vocabulary, http://www.w3.org/TR/void/
2. Bizer, C., Cyganiak, R., Heath, T.: How to publish Linked Data on the Web (2007), http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), Vol. 5(3), pp. 1--22 (2009)
4. Bosch, T., Wira-Alam, A., Mathiak, B.: Designing an Ontology for the Data Documentation Initiative. In: 8th Extended Semantic Web Conference (2011)
5. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF Data Cube vocabulary (2011), http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html
6. Data Documentation Initiative (DDI), http://ddialliance.org
7. DCMI Metadata Terms, http://dublincore.org/documents/2010/10/11/dcmi-terms/
8. FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/20100809.html
9. Gregory, A., Vardigan, M.: The Web of Linked Data. Realizing the Potential for the Social Sciences (2010), http://odaf.org/papers/201010_Gregory_Arofan_186.pdf
10. Harth, A.: VisiNav: A system for visual search and navigation on web data. J. Web Sem. 8(4), pp. 348--354 (2010)
11. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Proceedings of the 6th European Semantic Web Conference: Research and Applications (Heraklion, Crete, Greece) pp. 708--722 (2009)
12. King, G., Keohane, R., Verba, S.: Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton University Press (1994)
13. Kohler, U., Kreuter, F.: Datenanalyse mit STATA. Oldenbourg (2008)
14. Kruk, S.R., McDaniel, B.: Goals of Semantic Digital Libraries. In: Kruk, S.R., McDaniel, B. (eds.) Semantic Digital Libraries. Springer (2009)
15. Ladwig, G., Tran, D.T.: Linked Data Query Processing Strategies. In: Proceedings of the 9th International Semantic Web Conference (ISWC '10). Springer (2010)
16. Lazer, D., et al.: Computational Social Science, Science: 323 (5915), pp. 721--723 (2009)
17. Schnell, R., Hill, P., Esser, E.: Methoden der empirischen Sozialforschung. Oldenbourg (2005)
18. SKOS Simple Knowledge Organization System, http://www.w3.org/2004/02/skos/
19. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
20. Statistical data and metadata exchange (SDMX), http://sdmx.org/
21. Tran, D.T.: Semantic Web Search – A Process-Oriented Perspective on Data Retrieval on the Semantic Web (2010)
22. Wagner, A., Ladwig, G., Tran, D.T.: Browsing-oriented Semantic Faceted Search. In: Proc. of the 22nd Conf. on Database and Expert Systems Applications (DEXA). Springer (2011)
23. Zapilko, B., Mathiak, B.: Performing Statistical Methods on Linked Data. In: DC-2011: Proc. of the Int. Conference on Dublin Core and Metadata Applications, The Hague (2011)